DEÜ MÜHENDİSLİK FAKÜLTESİ
MÜHENDİSLİK BİLİMLERİ DERGİSİ
Cilt: 16 Sayı: 48 sh. 86-93 Eylül 2014

DESTEK VEKTÖR MAKİNELERİ PARAMETRE OPTİMİZASYONUNUN DUYGU ANALİZİ ÜZERİNDEKİ ETKİSİ

(EFFECTS OF SUPPORT VECTOR MACHINES PARAMETER OPTIMIZATION ON SENTIMENT ANAYLSIS)

Aysun GÜRAN[1], Mitat UYSAL[1], Özge DOĞRUSÖZ[2]

ÖZET/*ABSTRACT*

Kişilerin kullandıkları ürünler ve satın aldıkları hizmetler hakkındaki görüşlerini sosyal medya üzerinden paylaşması yorumların kategorize edilmesini sağlayan duygu analizi konusunun önem kazanmasını sağlamıştır. Duygu analizi ile ilgili çalışmalarda sınıflandırma metodu olarak destek vektör makineleri (DVM)'nin başarılı performansı pek çok kez vurgulanmıştır. Bu çalışma ile duygu analizinin gerçekleştirebileceği farklı veri setleri üzerinde DVM yöntem performansını etkileyen parametre değişimlerinin sınıflandırma performansı üzerindeki etkileri incelenmiş ve farklı deneyler sonucu elde edilen durumlar yorumlanmıştır.

Sentiment Analysis which has the meaning of categorization of comments has been popular since people share their ideas about the products and services that they bought. The studies about sentiment analysis point out the importance of support vector machines (SVM) many times. By this work, using different sentiment analysis data sets, parameter changes that effects the performance of SVM method have been analysed and different cases that are acquired by different experiments have been interpreted.

ANAHTAR KELİMELER/*KEYWORDS*

Duygu analizi, Destek vektör makineleri, Parametre optimizasyonu
Sentiment analysis, Support vector machines, Parameter optimization

[1] Doğuş Ün., Bilgisayar Müh., Böl., İSTANBUL, adogrusoz,muysal@dogus.edu.tr
[2] Doğuş Ün., İşletme., Böl., İSTANBUL, ozgedogrusoz@hotmail.com

1. GİRİŞ

İnternetteki veri miktarının genişlemesi ve sosyal medyanın ticarete olan etkisi firma-tüketici ilişkisini farklı bir boyuta taşımaktadır. İnsanların satın aldıkları ürün ve hizmetler hakkındaki olumlu olumsuz kişisel düşünceleri artık sadece firma ile tüketici arasında kalmamakta, sosyal medya üzerinden de paylaşılmaktadır. Yaşanan bu durumun etkisiyle sosyal medya ve bloglar aracılığı ile paylaşılan gönderilerin duygu analizi sürecinden geçirilerek otomatik olarak sınıflandırılması büyük bir önem taşımaktadır.

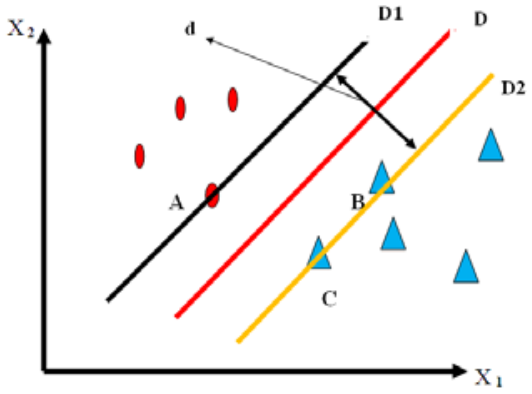
Literatürde Türkçe veri setlerinin kullanıldığı bir çok duygu analizi çalışması mevcuttur. Kayahan vd., çalışması ile seçilen TV programları için yapılan yorumları olumlu, olumsuz veya nötr olarak etiketlenmiş ve hesaplanan skor değerine bağlı olarak TV reyting sonuçlarının tutarlılığını test edilmiştir (Kayahan vd., 2013). Meral ve Diri, Twitter üzerinde Naïve Bayes (NB), Rastgele Orman (RO) ve DVM gibi makine öğrenmesi yöntemlerini kullanarak duygu analizini gerçekleştirmiş ve en başarılı sonuçlara DVM ile ulaşıldığını belirtmiştir. (Meral ve Diri, 2014; Çetin ve Amasyalı, 2013), eğiticili ve geleneksel terim ağırlıklandırma yöntemleriyle Türkçe Twitter gönderileri üzerinde duygu analizini gerçekleştirmiş ve yine NB, RO, DVM ve karar ağaçları (J48) arasından en başarılı sınıflandırma metodu olarak DVM'leri işaret etmiştir. DVM'lerin kullanıldığı bu çalışmalar, DVM'leri Weka yazılımında varsayılan parametreleri ile kullandıklarını belirtmişlerdir (Meral ve Diri, 2014; Çetin ve Amasyalı, 2013). Halbuki DVM'lerin kullanımında, yöntem performansını etkileyebilecek bir çok faktör mevcuttur. Bu faktörler, uygun çekirdek fonskiyonunun seçimi ve çekirdek fonksiyonuna ait uygun parametrelerin belirlenmesidir. Bu etmenler probleme göre uygun seçilmediği takdirde DVM'nin genelleştirme performansı olumsuz yönde etkilenecektir. Bu çalışma ile duygu analizinin uygulanabileceği üç veri seti üzerinde, DVM'lerde belirlenen çekirdek fonksiyonu için kullanılabilecek parametre değer değişimlerinin yöntem performansı üzerindeki etkisi araştırılmıştır. DVM kullanımında radyal tabanlı çekirdek fonksiyonu parametrelerini örgü arama (grid search) yöntemi ile belirleyen bir durum irdelenmiştir. Kullanılan veri setleri Kemik doğal dil işleme grubunun yayınlamış olduğu veri setleri içinden seçilmiştir. Sistem tasarımı için veri setlerine ait karakter tabanlı n-gramlar, kelime tabanlı n-gramlar, noktalama işaretleri gibi özellikleri içeren arff uzantılı dosyalar yaratılmış ve Weka LibSVM modülü kullanılarak sınıflandırma işlemi gerçekleştirilmiştir.

Makalenin geri kalan kısmı şu şekildedir: Makalenin ikinci bölümünde destek vektör makineleri anlatılmıştır; üçüncü bölümünde kullanılan veri setleri tanıtılmış ve sistem özelliklerinden söz edilmiştir. Sonuçlar bölümünde ise genel yorumlarda bulunularak gelecek çalışmalardan bahsedilmiştir.

2. DESTEK VEKTÖR MAKİNELERİ

Destek vektör makinesi iki boyutlu uzayda doğrusal, üç boyutlu uzayda düzlemsel ve çok boyutlu uzayda hiperdüzlem şeklindeki ayırma mekanizmaları ile veriyi iki ya da daha çok sınıfa ayırma yeteneğine sahiptir.

Veri grubunun bir doğru ile ayrılabildiği durum, grubun lineer olarak ayrılabildiği durumdur. Burada tarafından ileri sürülen bir fikir, iki sınıfı ayıran nesnenin bir doğru yerine bir koridor olması ve bu koridorun genişliğinin bazı veri vektörleri tarafından belirlenerek mümkün olan en büyük genişlikte olmasıdır (Cortes ve Vapnik, 1995). Şekil 1'de bu durum görülmektedir:



Şekil 1. İki sınıfın doğrusal olarak ayrılabildiği durum

D doğrusu <w.x>+b=0 denklemi ile belirlenmekte olup, w ağırlık vektörü ve b de sabit sayı (bias) değeridir. D1 doğrusu <w.x>+b=1 denklemi ile ve D2 doğrusu <w.x>+b=-1 denklemi ile belirlenir. D1 ve D2 doğruları arasındaki uzaklık (marjin) d ile gösterilirse, analitik geometri bilgileri ile, d'nin $(d = \frac{2}{||w||})$ şeklinde hesaplanabileceği kolayca görülür. ||w|| değeri $||w||=\sqrt{w1^2 + w2^2}$ şeklinde hesaplanır. d'nin maksimum değerinin elde edilmesi için doğal olarak ||w||'nin ya da $||w||^2$'nin minimum değerinin bulunması gerekir. Buna göre optimizasyon problemi,

$$\text{Min.} \quad \frac{1}{2}||w||^2 \tag{1}$$

Kısıtlar: (2)

$<w.x>+b \geq 1$ ise yi=1 (Sınıf I-Kırmızı)
$<w.x>+b \leq -1$ ise yi=-1 (Sınıf II-Mavi)

şeklinde bir kuadratik programlama problemi haline dönüşür. Burada iki kısıt aşağıdaki şekilde birleştirilebilir:

$$yi.(<w.x>+b) \geq 1 \tag{3}$$

Buna göre kuadratik programlama problem:

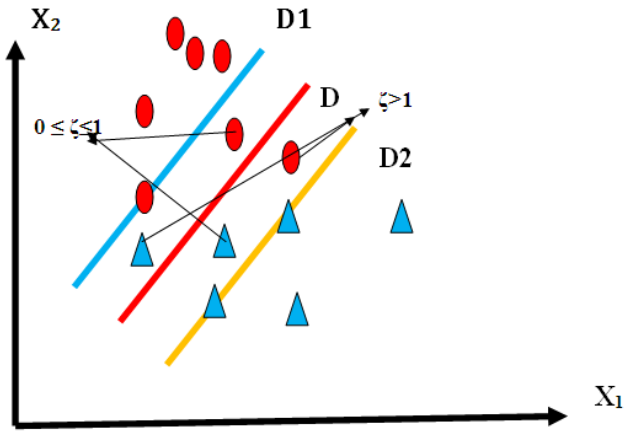
$$\text{Min.} \quad \frac{1}{2}||w||^2 \tag{4}$$

Kısıtlar:

$$yi.(<w.x>+b) \geq 1 \tag{5}$$

haline gelecektir. Bu problem primal problem olarak adlandırılır.

Şekil 2 ile iki sınıfın geniş bir marjin boyunca net olarak ayrılamadığı durum görülmektedir. Bu durumda sınırı aşma durumunu simgeleyen ζi parametresi amaç fonksiyonu içine girecek ve bu parametrenin de minimum olması istenecektir.



Şekil 2. İki sınıfın net olarak ayrılamadığı durum

Buna göre primal optimizasyon problem:

$$\text{Min. } J(w,\zeta) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\zeta_i \qquad (6)$$

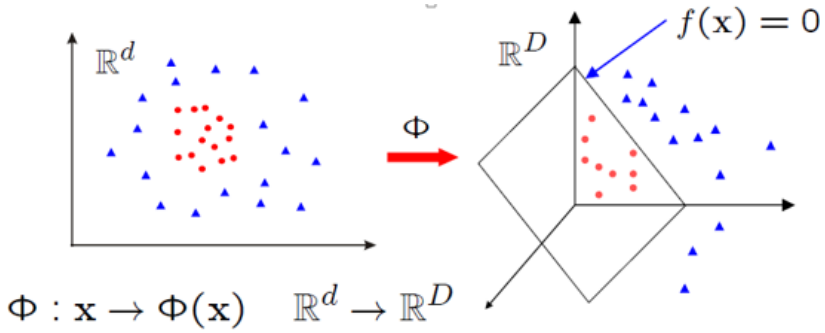
Kısıtlar:

$$y_i.(<w.x>+b) \geq 1-\zeta_i \quad i=1,2,...n \qquad (7)$$

$$\zeta_i \geq 0$$

şekline gelecektir. Burada ζi değişkenleri gevşek değişken olarak isimlendirilir. Burada,w,b ve ζi değişkenlerinin en iyi değerleri aranacaktır. C'nin değeri 0 ile sonsuz arasında herhangi bir değer olabilir. C parametresi modeli kuran kişi tarafından hatayı tolore etmek amacı ile uygun bir değer olarak seçilir. C değeri 0'a yaklaşırsa gevşek değişkenler kısıtsız hale gelir. Bu durumda marjin genişliği maksimum hale gelir (marjin içinde ya da yanlış tarafta kaç tane veri olduğu önemini yitirir). Bu durum bazı hallerde yararlıdır. Öğrenme fazında hassasiyet azalır ancak yeni data kabulünde başarı artar. C'nin sonsuza yaklaşması hassasiyeti çok arttırır (marjin genişliği çok azalır). Öğrenme fazında hassasiyet yüksek ancak yeni veri kabulünde güçlükler oluşur. C'nin değeri kullanıcının tecrübesine göre deneme-yanılma yöntemi ile belirlenir.

Birçok veri kümesi için, iki boyutlu durumdaki veriler, doğrusal ayraç yardımı ile birbirinden ayrılabilir değildir. Şekil 3'te bu durum görülmektedir:



Şekil 3. Verilerin lineer bir ayraç ile ayrılamadığı durum

$x \rightarrow \Phi(x)$ dönüşümünün yapılması ile ayırıcı düzlem $f(x) = w.\Phi(x) + b = 0$ haline gelecektir. İki boyutlu uzaydaki ifadelerde, $<x_i.x_j>$ skaler çarpımları yerine çok boyutlu uzayda $<\Phi i.\Phi_j>$ şeklindeki iç çarpımlar gelecektir. Bu durumda $<\Phi i.\Phi j>$ çarpımı yerine, $K(X_1,X_2) = \Phi(X_1).\Phi(X_2) = \Phi(X_1)^T.\Phi(X_2) = <\Phi(X_1),\Phi(X_2)>$ şeklinde tanımlanan bir K fonksiyonu kullanılabilirse o takdirde özellikle çok büyük veri miktarları için çok büyük zaman harcanarak hesaplanabilecek olan $<\Phi(X_1),\Phi(X_2)>$ iç çarpımları yerine $K(X_1,X_2)$ ifadesi ile bu hesaplamaları çok daha az işlem ve zaman harcayarak gerçekleştirmek mümkün olacaktır. Burada K fonksiyonu bir çekirdek (kernel) adını alır; $<\Phi(X_1),\Phi(X_2)>$ ifadesi yerine K'nın kullanılması ise çekirdek püf noktası (kernel trick) olarak isimlendirilir. Çizelge 1'de, K kernel fonksiyonu için kullanılabilecek alternatifler hakkında bilgi verilmektedir:

Çizelge 1. Çekirdek fonksiyonları

Çekirdek Adı	Formülü	Parametreleri
Doğrusal	x^Ty+c	yok
Polinomyal (Polinomial)	$(x^Ty+1)^d$	d
Radyal tabanlı (Radial Basis)	$\exp(-\gamma\|x-y\|^2)$	γ
Sinir ağı sigmoid (Neural network sigmoid)	$\tanh(a\,x^Ty+b)$	a,b

DVM'lerde dikkate alınması gereken önemli bir konu da büyük veri gruplarının belirli özelliklere göre ikiden fazla gruba ayrılma durumudur. Destek vektör makinesini çok sayıda sınıf durumunda kullanabilmek için, problem çok sayıda ikili sınıf problemine dönüştürülmelidir. En çok kullanılan yaklaşımlar: Biri ve diğerleri (One vs All) yaklaşımı ile Bire bir (One vs One) yaklaşımıdır (Hsu ve Lin, 2002). Bu çalışmada bire bir yaklaşımı kullanılmıştır.

3. KULLANILAN VERİ SETLERİ VE SİSTEM ÖZELLİKLERİ

Bu çalışmada Kemik doğal dil işleme grubunun yayınlamış olduğu toplam üç veri seti kullanılmıştır. VS1, VS2, VS3 sırasıyla birinci, ikinci ve üçüncü veri setini ifade etmek üzere, bu veri setleri ile ilgili bilgiler Çizelge 2 ile belirtilmiştir:

Çizelge 2. Veri setleri tanıtımı

Veri Seti	Açıklama
VS1	Bu set 3 farklı sınıfa ait (olumlu, olumsuz, nötr) 3000 adet tweet içermektedir.
VS2	Bu set blog yazarlarının 4 farklı ruh haliyle (neşeli, sinirli, üzgün, karışık) yazdıkları 157 adet blog yazısını içermektedir.
VS3	Bu set 3 farklı yorum sınıfına (olumlu, olumsuz, nötr) ait 105 adet film yorumu içermektedir.

DVM'lerin kullanımında, yöntemin performansını etkileyebilecek faktörler, uygun çekirdek fonksiyonunun seçimi ve çekirdek fonksiyonuna ait uygun parametrelerin belirlenmesidir. Bu faktörler probleme göre uygun seçilmediği takdirde, DVM'nin genelleştirme performansı olumsuz yönde etkilenecektir. Bu makalede, DVM yöntemine ait radyal tabanlı çekirdek fonksiyonu parametresini (γ) ve C ceza parametresini örgü arama (grid

search) yöntemi ile belirleyen bir durum irdelenmiştir. Örgü arama yöntemi, 10 kat çapraz geçerleme (10-fold cross validation) ile, ilgilenilen parametre uzayını uygun aralıklarda ayrıklaştırmakta ve her bir düğüm noktasına karşılık gelen değerleri çekirdek fonksiyonu parametreleri olarak DVM yöntemine girdi olarak vermektedir (Kaya ve Kaya, 2014). Ardından geçerleme verileri üzerinde en yüksek sınıflandırma performansını veren parametre değerleri en uygun parametre olarak seçilmektedir. C ve γ değerleri örgü arama metoduna göre $C=\{10^1,10^2, 10^3,10^4\}$, $\gamma=\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ kümelerinin kartezyen çarpımı ile elde edilen ikililer kullanılarak belirlenmiştir. DVM yönteminin belirtilen şartlarda uygulanması için Weka LibSVM modülü kullanılmıştır.

4. SONUÇLAR VE GELECEK ÇALIŞMALAR

Bu bölümde DVM yöntem performansının farklı γ ve C parametrelerine göre incelenen veri setleri üzerindeki başarım değerleri incelenmiştir. Çizelge 3, parametrelerin VS1 üzerindeki etkisini göstermektedir. Burada veri seti hazırlanırken özellik olarak karakter tabanlı 2-gram, 3-gram, 4-gram ve 5-gramlar birlikte kullanılmıştır. Kullanılan özellikler veri seti referansı ile indirildiğinde elde edilen dosyalar arasından "texts_in_arff" dosyasından görülebilir (Kemik doğal dil işleme grubu, 2014). Weka yazılımının girdisi olan arff dosyası n-gram özelliklerinin baz alınmasıyla kelime sıklığı (term frequency) metriği kullanılarak oluşturulmuştur. Bu şartlar altında toplam 15368 adet özellik elde edilmiştir. Bu dosya üzerinde Korelasyon Tabanlı Özellik Seçici (Correlation-based Feature Selection-CFS) özellik seçim algoritması uygulanarak özellik sayısı 247'e düşürüldükten sonra yapılan farklı deneyler ile Çizelge 3'teki başarım değerleri elde edilmiştir.

Çizelge 3. Parametre seçiminin VS1 üzerindeki etkisi

	$C:10^4$	$C:10^3$	$C:10^2$	$C:10^1$
γ: 10^{-6}	0,556	0,452	**0,429**	0,429
γ: 10^{-5}	0,622	0,557	0,452	0,429
γ: 10^{-4}	**0,633**	0,622	0,557	0,452
γ: 10^{-3}	0,624	0,633	0,621	0,56
γ: 10^{-2}	0,558	0,594	0,624	0,622
γ: 10^{-1}	0,543	0,542	0,548	0,594

Çizelge 3 incelendiğinde parametre değişikliklerinin DVM performansı üzerinde oldukça etkili olduğu görülmektedir. En düşük performans (0,429) ile en başarılı performans (0,633) değeri arasında oldukça yüksek bir farklılık gözlemlenmiştir.

Çizelge 4, incelenen farklı parametrelerin VS2 üzerindeki etkisini göstermektedir. Burada veri seti hazırlanırken özellik olarak karakter 2-gram, 3-gram, kelime kökleri, noktalama işaretlerinin sıklığı gibi farklı özellikler kullanılmıştır. Kullanılan özellikler veri setinin referansından indirildikten sonra görülebilecek olan "ozellikler.txt" dosyasından incelenebilir (Kemik doğal dil işleme grubu, 2014). VS2'de incelenen toplam özellik sayısı 23018 adettir. CFS özellik seçim algoritmasından sonra bu sayı 54'e düşürülmüştür. Çizelge 4'de belirtilen DVM başarım değerleri bu şartlar altında oluşturulmuştur. Çizelgeden görüldüğü üzere en yüksek başarım değeri 0,751 aynı veri seti üzerinde farklı parametreler ile elde edilen en düşük başarım oranı ise 0,305'dir. Başarım oranı farkı oldukça yüksektir.

Çizelge 4. Parametre seçiminin VS2 üzerindeki etkisi

	C:10^4	C:10^3	C:10^2	C:10^1
γ: 10^{-6}	0,694	0,643	0,318	**0,305**
γ: 10^{-5}	**0,751**	0,70	0,643	0,318
γ: 10^{-4}	0,713	0,738	0,687	0,624
γ: 10^{-3}	0,694	0,694	0,719	0,713
γ: 10^{-2}	0,707	0,707	0,707	0,719
γ: 10^{-1}	0,528	0,528	0,528	0,528

Çizelge 5, parametrelerin VS3 üzerindeki etkisini göstermektedir. Burada veri seti hazırlanırken özellik olarak VS2'de olduğu gibi karakter 2-gram, 3-gram, kelime kökleri, noktalama işaretlerinin sıklığı gibi farklı özellikler kullanılmıştır. Kullanılan özellikler yine veri setinin referansından indirildikten sonra görülebilecek olan "ozellikler.txt" dosyasından incelenebilir (Kemik doğal dil işleme grubu, 2014). VS3'de incelenen toplam özellik sayısı 6439 adettir. CFS özellik seçim algoritmasından sonra bu sayı 64'e düşürülmüştür. Sonuçlara göre en yüksek başarım oranı 0,752 iken en düşük başarım oranı yüksek bir düşüş ile 0,342 olarak elde edilmiştir.

Çizelge 5. Parametre seçiminin VS3 üzerindeki etkisi

	C:10^4	C:10^3	C:10^2	C:10^1
γ: 10^{-6}	0,514	0,371	0,342	0,342
γ: 10^{-5}	**0,752**	0,514	0,371	0,342
γ: 10^{-4}	0,752	0,752	0,514	0,371
γ: 10^{-3}	0,658	0,752	0,752	0,523
γ: 10^{-2}	0,714	0,714	0,752	0,704
γ: 10^{-1}	0,666	0,666	0,666	0,666

Bu deneylerin sonucunda seçilen farklı parametrelerin DVM başarımı üzerinde etkili olduğu gözler önüne sergilenmiştir. Bu durum incelenemeyen diğer aralıklar için çok daha yüksek başarım sonuçlarını karşımıza çıkarabilir. Üstelik farklı çekirdek fonksiyonları da birbirinden farklı sonuçlar sergileyecektir. Dolayısıyla DVM'ler uygulanırken parametre optimizasyonu için hızlı metotların önerilmesi sistemin çalışma süresi ve performansı adına çok önemlidir. Bundan sonraki çalışmalarımızda amacımız parametre optimizasyonu için sezgisel metotlar üzerinde durmak olacaktır.

KAYNAKLAR

Kayahan D., Sergin A., Diri B. (2013): "Twitter ile TV Program Reytinglerinin Belirlenmesi", IEEE 21. Sinyal İşleme ve İletişim Uygulamaları Kurultayı, SIU 2013, Kıbrıs.

Meral M., Diri B. (2014): "Twitter Üzerinden Duygu Analizi", IEEE 22. Sinyal İşleme ve İletişim Uygulamaları Kurultayı, SIU 2014, Trabzon.

Çetin M., Amasyalı F. (2013): "Eğiticili ve Geleneksel Terim Ağırlıklandırma Yöntemleriyle Duygu Analizi", IEEE 21.Sinyal İşleme ve İletişim Uygulamaları Kurultayı, SIU 2013, Kıbrıs.

Cortes C., Vapnik V. (1995): "Support Vector Networks", Machine Learning, Cilt 20, s.273–297.

Hsu C. W., Lin C. J. (2002): "A Comparison of Methods for Multiclass Support Vector Machines", IEEE Transactions On Neural Networks, Cilt 13, No. 2, s.415-425.

Kemik Doğal Dil İşleme Grubu (2014): http://www.kemik.yildiz.edu.tr/.

Kaya G. T., Kaya H. (2014): "Destek Vektör Makinaları Model Parametrelerinin Yüksek Boyutlu Model Gösterilimi ile Optimizasyonu ve Hiperspektral Görüntülere Uygulanması", IEEE 22. Sinyal İşleme ve İletişim Uygulamaları Kurultayı, SIU 2014, Trabzon.